Evaluating Personal Information Retrieval

Liadh Kelly, Paul Bunbury, and Gareth J. F. Jones

Centre for Next Generation Localisation,
School of Computing, Dublin City University (DCU), Dublin 9, Ireland
{lkelly,gjones}@computing.dcu.ie

Abstract. Evaluation of personal search over an individual's personal information space on the desktop or elsewhere is problematic for reasons relating both to the personal and private nature of the data and the associated personal information needs of collection owners. Indeed challenges associated with evaluation in this space are recognised as one of the key factors hindering the development of research in personal information retrieval. We present the "personal information retrieval evaluation (PIRE)" tool, which provides a solution to this evaluation problem using a 'living laboratory' approach. This tool allows for the evaluation of retrieval techniques using 'real' individuals' personal collections, queries and result sets, in a cross-comparable repeatable way, while importantly maintaining an individual's informational privacy.

Key words: Living lab, personal information retrieval evaluation.

1 Introduction

The full value of personal information archives, such as those found on the desktop, can only be realised if they can be searched effectively. Development of suitable information retrieval (IR) technologies for this personal space requires that their effectiveness be evaluated through measurement of retrieval accuracy. A standard IR evaluation collection includes a document collection, a test set of information needs expressed as search topics, and a set of judgments indicating the relevance of documents to each test topic. However, for personal search, such a dataset is difficult to generate due to the heterogeneous nature of personal information spaces, practical challenges of collecting the data, and significantly, privacy concerns relating to the personal nature of this data. This latter issue creates problems for all aspects of the evaluation of search of personal collections. Current work on evaluation of personal collection search is exploring the development of simulated personal Cranfield type search test collections [2]. However, these collections do not represent the diversity of real users collections, items selected by an individual owning the collection that they actually want to retrieve from it, nor the query terms collection owners will use. From the search perspective, the key difference between search on personal collections and standard search environments, is that only the owner of the personal collection will be aware of the contents, and thus only they will be able to establish information

needs which can be answered by the collection and to determine the relevance of returned content. In order to satisfy this requirement, real users are needed to perform test search tasks, preferably on their own personal data. This requires not only that a user participates in evaluation experiments, but also that they enable the archiving of their personal data and for it to be processed for use in a search system. To date, in order to conduct experiments in this space, researchers have largely needed to create their own test collections consisting of individuals' data, queries and result sets (e.g. [4]). There are three problems with this approach: 1) the effort required to create these collections; 2) the difficulty in gaining large volumes of subjects for such experiments; and 3) the lack of cross-comparability across research institutions working with different subjects. In this paper we present a strategy to support evaluation of personal collection search in a cross-comparable way based on real user data.

2 Living Laboratory Evaluation Framework

Our personal search evaluation methodology presented here is similar to the concept of the living laboratory described in [1]. This was proposed in the context of evaluating information-seeking support systems which aim to assist users in carrying out open-ended search related tasks. The basic idea of the living laboratory is that rather than individual research groups independently developing experimental search infrastructures and gathering their own groups of test searchers, that an experimental environment is developed which facilitates sharing of resources. We propose that within a living laboratory for personal search evaluation researchers wishing to evaluate their technologies would participate in a collaborative evaluation effort. Common indexing and search components would be made available to individuals who agree to take part in the evaluation exercise. This would then be used to gather personal collections locally and conduct search experiments. In the next section we present our prototype tool developed for this evaluation paradigm.

3 Personal Information Retrieval Evaluation (PIRE) Tool

The PIRE tool (see Figure 1) is a cross platform application, developed in Java using the open source Terrier[1] indexing and search library. The PIRE tool runs on an individual's own machine and provides a means for them to index their personal desktop collection. Individuals can then perform queries on their indexed collection and then indicate retrieved items which are relevant to their queries. Through this process the individual creates a personal 'Cranfield' style test collection. Using the individual's personal test collection, the tool allows for the evaluation of retrieval approaches and generates a computed evaluation metrics file for the investigator. The remainder of this section describes the PIRE tool, and methodology underlying it in greater detail.

[1] http://www.terrier.org

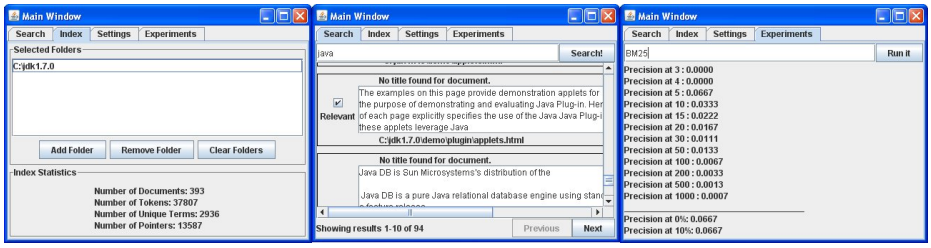

Fig. 1. The PIRE tool. From left to right: (a) indexing facility; (b) query entry and relevance judgement; (c) retrieval algorithm evaluation.

3.1 Data Collection: Indexing

The index tab (Figure 1(a)) is used by the subject to indicate the desktop folders to index and initiate index generation. For each item indexed the static index consists of item content, title and extension type. Currently supported file formats are: plain text files (.txt, .bib, .java, etc); PDF files; HTML and XML files; and files in the OOXML format (which includes XLSX, DOCX and PPTX files) and OLE2 format (which includes XLS, DOC and PPT files).

3.2 Data Collection: Queries and Result Sets

In order to evaluate retrieval approaches, queries and corresponding relevant result sets are required. The search tab of the PIRE tool (Figure 1(b)) allows the subject to enter queries and indicate the relevant retrieved items from their personal collection. Queries and relevance judgements are saved by the tool for future use by the evaluation component, described in the next section.

Query Generation Search topics within a standard IR test collection are typically pre-defined written statements known to be covered in the collection. Since the specific details of a personal collection will vary and not be known to the experimenter, broader search tasks would need to be defined for query generation using our system, e.g. referring to meetings with unnamed friends. The searcher would then generate a specific query for their collection from the general topic statement. Of course, even with these more general task statements, the searcher may sometimes find that they are unable to recall any relevant content. How to develop suitable task descriptions will obviously have to be clearly defined, and useful lessons in doing this may be gathered from work in designing less specific exploratory search tasks for evaluating information-seeking support systems [1].

Result Set Generation In order to gather relevance data for personal search, the PIRE tool enables the searcher to assess the relevance of items retrieved in response to each of their search tasks. Assessing all items in a collection for relevance is impractical. However, assessing only the items retrieved at high rank using one retrieval method may not give a reasonable indication of the

available relevant documents in the collection. To address this issue, pooling of results from runs using multiple retrieval methods is used in PIRE by applying the users search to multiple retrieval algorithms. The results of the pooling are shown to the searcher for assessment. The title, file path and abstract of items for assessment are presented to the searcher (Figure 1(b)). The original item can be obtained by clicking on the file path. Subjects indicate relevant items using the 'relevant' check box.

3.3 Evaluation

Using the generated user data sets, queries and relevant result sets, existing and new IR algorithms, such as those described in [3], can be tested using the experiments tab (Figure 1(c)) of the PIRE tool. Source code for algorithms to investigate can be distributed to experimental subjects by participating institutions, and simply loaded into the search system using the PIRE tool. The tool then runs the queries in the subject's collection using the IR algorithm to be tested, and then uses the subject's corresponding results set to generate a file containing performance measures, which is returned to the investigator. Generated performance measures are currently created via Terrier's `trec_eval` component, and consist of metrics such as precision, recall and various averages.

4 Conclusions

This paper presented the PIRE tool for personal information search evaluation using a living laboratory approach. The scenario is based on users maintaining their own personal collection on their own computer, and using standardized tools to index and search their collection. All relevance assessment and evaluation is carried out on their computer using the PIRE tool with only the computed evaluation metrics being returned for aggregation, using a suitable meta-analysis approach, thus preserving privacy of experimental subjects' personal data. Future versions of the tool will incorporate other item types (e.g. emails) and integrate a computer activity logging component, hence making it possible to log web activity and tag personal items with richer context sources such as 'date-time' of previous access related information.

Acknowledgments. This work was supported by DCU School of Computing.

References

1. D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation Challenges and Direction for Information-Seeking Support Systems. *IEEE Computer*, 42(3):60–66, 2009.
2. J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *Proceedings of CIKM 2009*, pages 1297–1306. 2009.
3. C. Manning, P. Raghavan, and H. Schutze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.
4. C. A. N. Soules and G. R. Ganger. Connections: Using Context to Enhance File Search. In *Proceedings of SOSP 2005*, pages 119–132. 2005.